Exploratory Data Analysis for Automobile Dataset

Purpose and Scope of The EDA

The purpose of the Exploratory Data Analysis for the Automobile dataset is to understand the fuel efficiency of the cars in the dataset and identify the factors that affect fuel efficiency. The scope of the EDA is to analyze the following columns: Fuel Type, city mpg, highway mpg, engine size, price, make, and body styles.

Through visualization and statistical analysis, the EDA aims to identify any trends or patterns in fuel efficiency and determine the impact of various features on fuel efficiency. The analysis can help in making decisions regarding car purchases, fuel consumption, and identifying the most fuel-efficient cars. Additionally, it can help car manufacturers in designing and producing more fuel-efficient cars by understanding the factors that affect fuel efficiency.

We will be answering the following questions:

1. Which car manufacturers produce the most fuel-efficient cars?
2. How does the engine size affect fuel efficiency?
3. Is there a difference in fuel efficiency based on the type of fuel used (gas or diesel)?
4. How does the body style of a car impact its fuel efficiency?
5. Does the price of a car have any relationship with its fuel efficiency?

In [1]:

```
# Import Libraries
import pandas as pd
import numpy as np
from scipy import stats
from mlxtend.preprocessing import minmax_scaling
import seaborn as sns
import missingno
import matplotlib.pyplot as plt
import networkx as nx
from mpl_toolkits.mplot3d import Axes3D
```

In [2]:

```
df = pd.read_csv('automobile.txt')
df
```

Out[2]:

	symboling	normalized-losses	make	fuel-type	aspiration	num-of-doors	body-style	drive-wheels	engine-location
0	3	?	alfa-romero	gas	std	two	convertible	rwd	front
1	3	?	alfa-romero	gas	std	two	convertible	rwd	front
2	1	?	alfa-romero	gas	std	two	hatchback	rwd	front
3	2	164	audi	gas	std	four	sedan	fwd	front
4	2	164	audi	gas	std	four	sedan	4wd	front
...	...	...	...	...	...	...	...	...	...
200	-1	95	volvo	gas	std	four	sedan	rwd	front
201	-1	95	volvo	gas	turbo	four	sedan	rwd	front
202	-1	95	volvo	gas	std	four	sedan	rwd	front
203	-1	95	volvo	diesel	turbo	four	sedan	rwd	front
204	-1	95	volvo	gas	turbo	four	sedan	rwd	front

205 rows × 26 columns

In [3]:

```
# Check the columns in Dataset so that we know the type of Data we can expect
df.columns
```

Out[3]:

```
Index(['symboling', 'normalized-losses', 'make', 'fuel-type', 'aspiratio
n',
       'num-of-doors', 'body-style', 'drive-wheels', 'engine-location',
       'wheel-base', 'length', 'width', 'height', 'curb-weight', 'engine-
type',
       'num-of-cylinders', 'engine-size', 'fuel-system', 'bore', 'strok
e',
       'compression-ratio', 'horsepower', 'peak-rpm', 'city-mpg',
       'highway-mpg', 'price'],
      dtype='object')
```

In [4]:

```
# Display the summarry of the Dataset
df.describe()
```

Out[4]:

	symboling	wheel-base	length	width	height	curb-weight	engine-size
count	205.000000	205.000000	205.000000	205.000000	205.000000	205.000000	205.000000
mean	0.834146	98.756585	174.049268	65.907805	53.724878	2555.565854	126.907317
std	1.245307	6.021776	12.337289	2.145204	2.443522	520.680204	41.642693
min	-2.000000	86.600000	141.100000	60.300000	47.800000	1488.000000	61.000000
25%	0.000000	94.500000	166.300000	64.100000	52.000000	2145.000000	97.000000
50%	1.000000	97.000000	173.200000	65.500000	54.100000	2414.000000	120.000000
75%	2.000000	102.400000	183.100000	66.900000	55.500000	2935.000000	141.000000
max	3.000000	120.900000	208.100000	72.300000	59.800000	4066.000000	326.000000

Data Cleaning and Preparation

1.Check for missing values:

we will Check if there are any missing values in the dataset and decide whether to drop or fill them.

In [5]:

```
# get the number of missing data points per column
missing_values_count = df.isnull().sum()

# look at the # of missing points in the first ten columns
missing_values_count[0:26]
```

```
num-of-doors         0
body-style           0
drive-wheels         0
engine-location      0
wheel-base           0
length               0
width                0
height               0
curb-weight          0
engine-type          0
num-of-cylinders     0
engine-size          0
fuel-system          0
bore                 0
stroke               0
compression-ratio    0
horsepower           0
peak-rpm             0
city-mpg             0
highway-mpg          0
```

we find out that the are no missing values, however we find that the is ? that exist in the price column so we will replace it and make the column numerical

In [6]:

```
# replace "?" with NaN
df.replace("?", float("nan"), inplace=True)

# convert price column to numeric data type
df["price"] = pd.to_numeric(df["price"])

# drop rows with NaN values in price column
df.dropna(subset=["price"], inplace=True)

# print the min and max price
print("Minimum price:", df["price"].min())
print("Maximum price:", df["price"].max())
```

```
Minimum price: 5118.0
Maximum price: 45400.0
```

2.Check for duplicates:

We can check if there are any duplicate rows in the dataset and remove them.

In [7]:

```
# check for duplicates
duplicate_rows = df.duplicated()
print(duplicate_rows)

# count the number of duplicates
print(duplicate_rows.sum())

# remove duplicates
df.drop_duplicates(inplace=True)
```

```
0      False
1      False
2      False
3      False
4      False
       ...
200    False
201    False
202    False
203    False
204    False
Length: 201, dtype: bool
0
```

We find out that the are no duplicates

3. Remove unnecessary columns:

We will then Remove Unecessary columns that will not be relavant to our analysis

In [8]:

```
# We find out that the is ? in the normalized-losses column
# We will drop this column as it is irrelavant for our analysis
#df.drop(["normalized-losses"],axis = 1, inplace = True)
df.head()
```

Out[8]:

	symboling	normalized-losses	make	fuel-type	aspiration	num-of-doors	body-style	drive-wheels	engine-location	wh b
0	3	NaN	alfa-romero	gas	std	two	convertible	rwd	front	
1	3	NaN	alfa-romero	gas	std	two	convertible	rwd	front	
2	1	NaN	alfa-romero	gas	std	two	hatchback	rwd	front	
3	2	164	audi	gas	std	four	sedan	fwd	front	
4	2	164	audi	gas	std	four	sedan	4wd	front	

5 rows × 26 columns

In [9]:

```
df.columns
```

Out[9]:

```
Index(['symboling', 'normalized-losses', 'make', 'fuel-type', 'aspiratio
n',
       'num-of-doors', 'body-style', 'drive-wheels', 'engine-location',
       'wheel-base', 'length', 'width', 'height', 'curb-weight', 'engine-
type',
       'num-of-cylinders', 'engine-size', 'fuel-system', 'bore', 'strok
e',
       'compression-ratio', 'horsepower', 'peak-rpm', 'city-mpg',
       'highway-mpg', 'price'],
      dtype='object')
```

In [10]:

```
# Drop the following columns
df.drop(["symboling","normalized-losses", "aspiration","num-of-doors","drive-wheels","er
```

Data Analysis and Visualisation

In [11]:

```
df
```

Out[11]:

	make	fuel-type	body-style	engine-size	city-mpg	highway-mpg	price
0	alfa-romero	gas	convertible	130	21	27	13495.0
1	alfa-romero	gas	convertible	130	21	27	16500.0
2	alfa-romero	gas	hatchback	152	19	26	16500.0
3	audi	gas	sedan	109	24	30	13950.0
4	audi	gas	sedan	136	18	22	17450.0
...	...	...	...	...	...	...	...
200	volvo	gas	sedan	141	23	28	16845.0
201	volvo	gas	sedan	141	19	25	19045.0
202	volvo	gas	sedan	173	18	23	21485.0
203	volvo	diesel	sedan	145	26	27	22470.0
204	volvo	gas	sedan	141	19	25	22625.0

201 rows × 7 columns

1. Univeriate Analysis

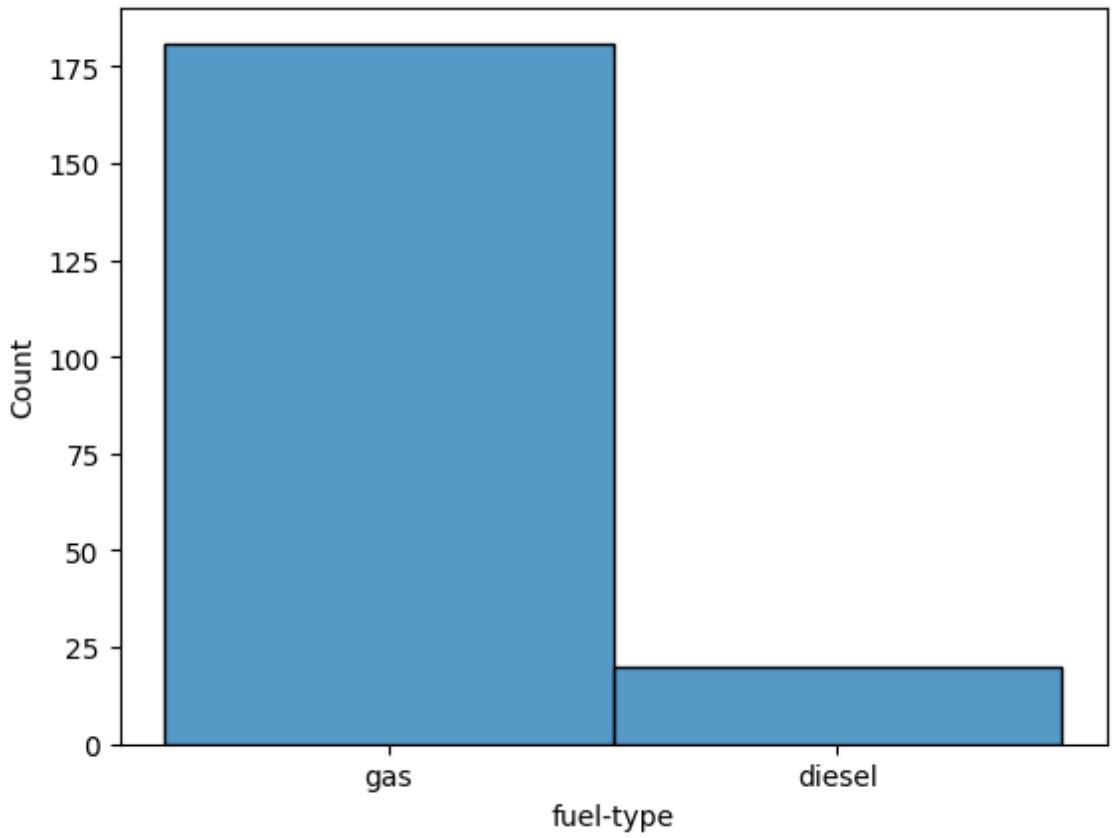
```
we will now analyse the data by Examining only one variable at a time
```

```
Distribution of Fuel Types:

- This will help us understand the proportion of each type of fuel in the dataset, we will then know which fuel type is dominant between gas and diesel.
```

In [12]:

```
sns.histplot(data=df, x='fuel-type')
plt.show()
```



Findings:

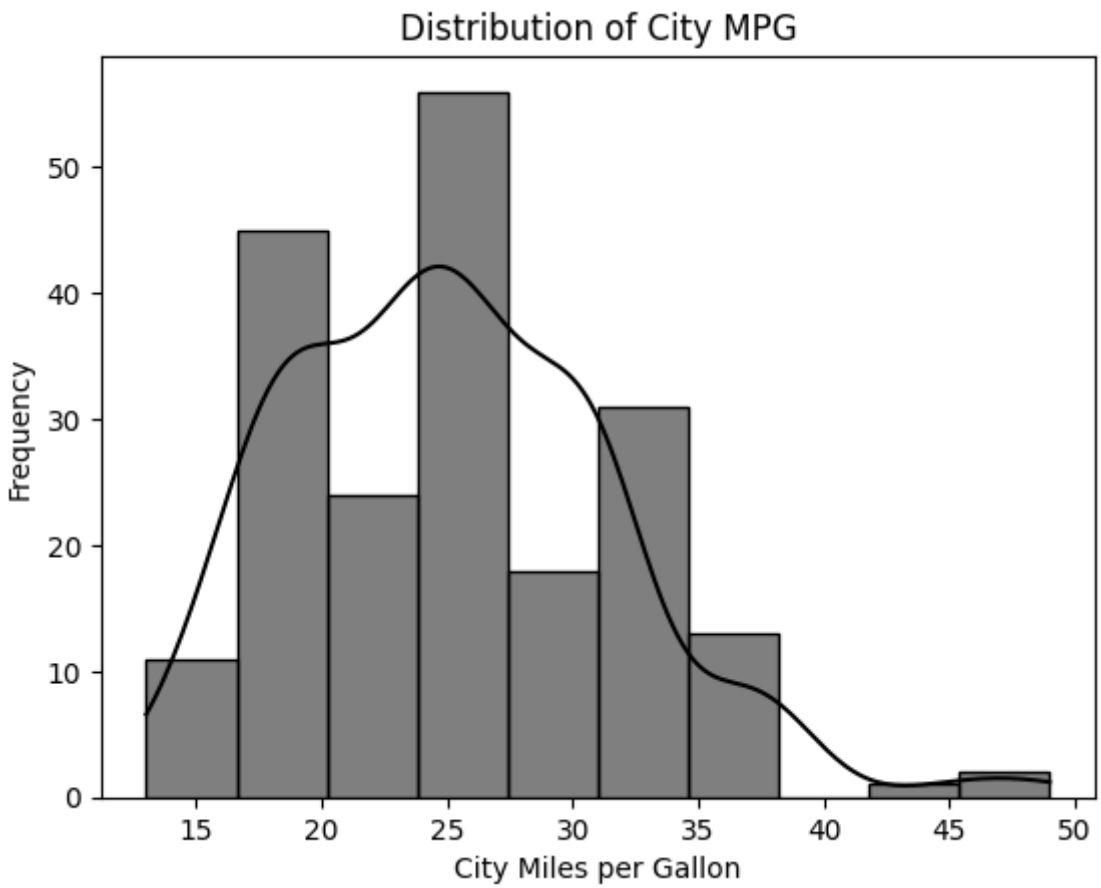
Majority of the cars in the Dataset use gas rather than diesel.

Distribution of City MPG:

The purpose is to see if most of the cars in the dataset are fuel efficient when driving in city traffic, this will be done by visualising the distribution of the values using a histogram and Boxplot.

In [13]:

```
# create histogram for city-mpg
sns.histplot(data=df, x='city-mpg', kde=True, color="black")
plt.title('Distribution of City MPG')
plt.xlabel('City Miles per Gallon')
plt.ylabel('Frequency')
plt.show()
```

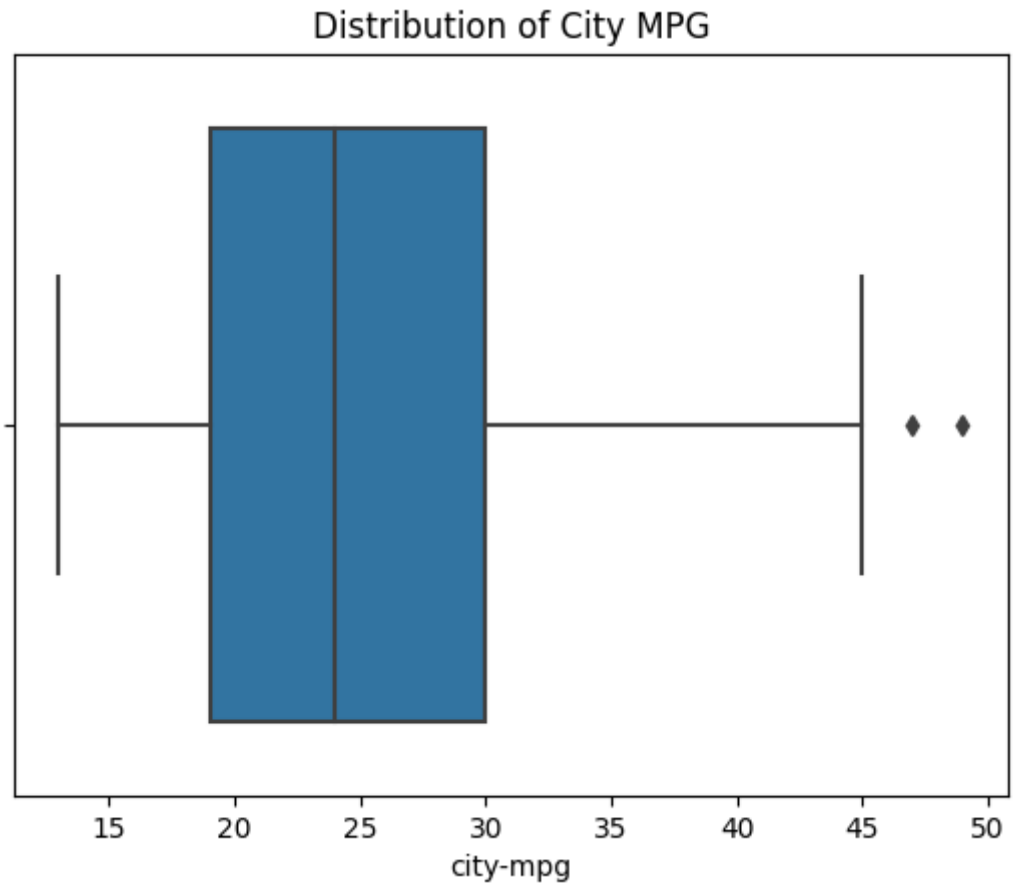


In [14]:

```
# Plot a Boxplot for City-mpg
sns.boxplot(data=df, x='city-mpg')
plt.title('Distribution of City MPG')
```

Out[14]:

```
Text(0.5, 1.0, 'Distribution of City MPG')
```



Findings:

- We find out that most of the cars have a score of 25, indicating they are above average.

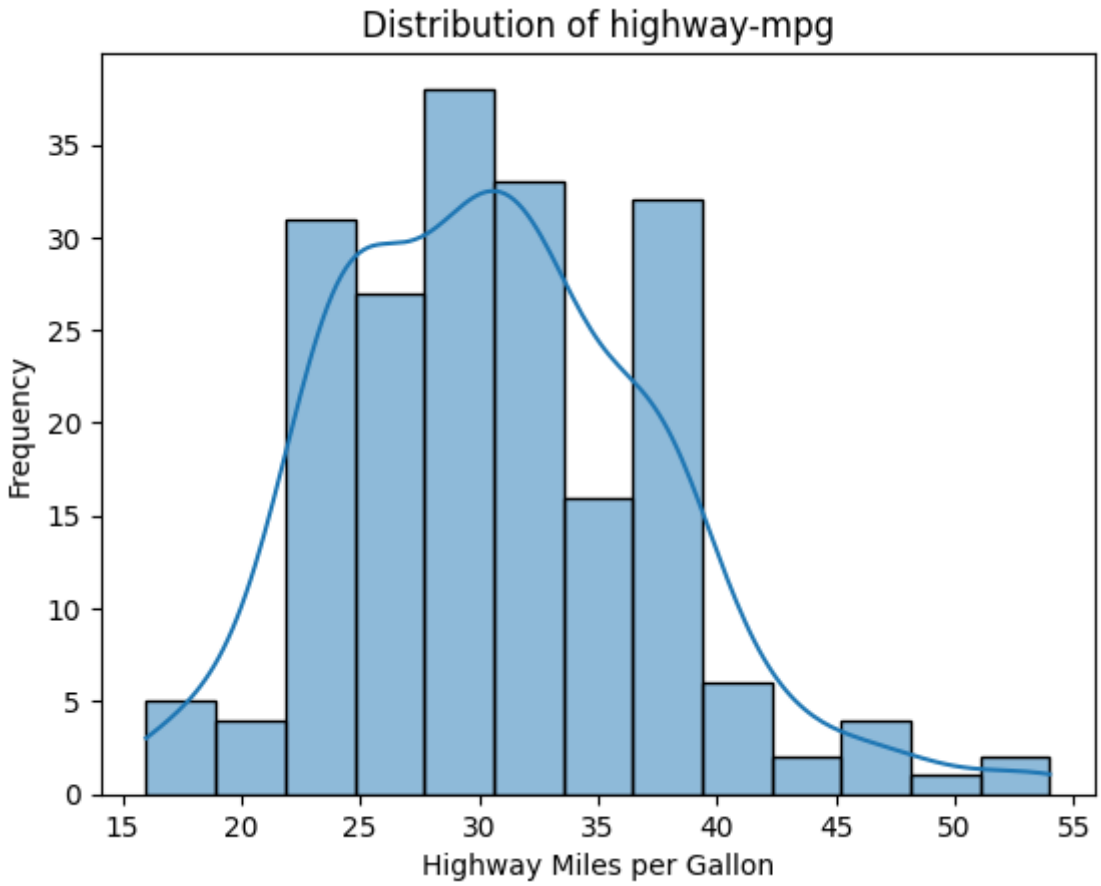
- This suggests majority of the cars are fuel-efficient while driving in city traffic, this is supported by the positively skewed Boxplot and distribution line.

Distribution of Highway MPG:

The purpose is to see if most of the cars in the dataset are fuel efficient while driving on a highway , this will be done by visualising the distribution of the values using a histogram and Boxplot.

In [15]:

```
# create histogram for highway-mpg
sns.histplot(data=df, x='highway-mpg', kde=True)
plt.title('Distribution of highway-mpg')
plt.xlabel('Highway Miles per Gallon')
plt.ylabel('Frequency')
plt.show()
```

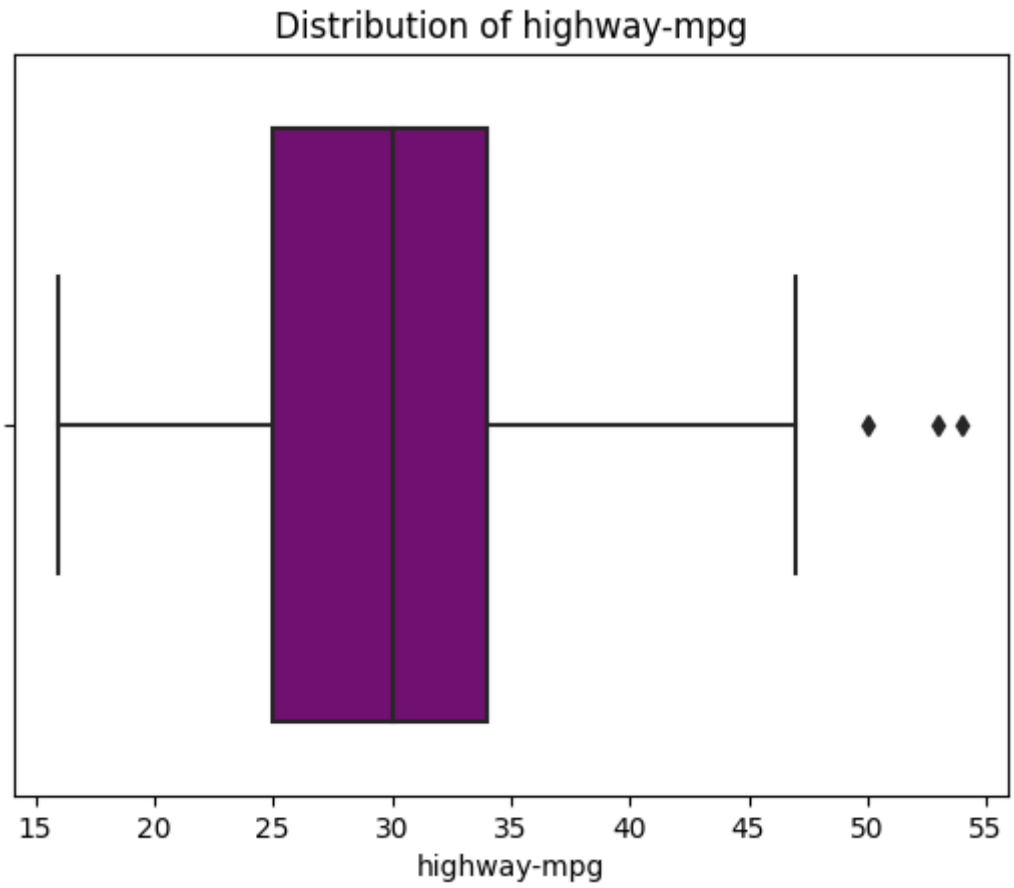


In [16]:

```
# Plot a Boxplot for highway-mpg
sns.boxplot(data=df, x='highway-mpg', color='purple')
plt.title('Distribution of highway-mpg')
```

Out[16]:

```
Text(0.5, 1.0, 'Distribution of highway-mpg')
```



Findings:

- We find out that most of the cars have a score below average.

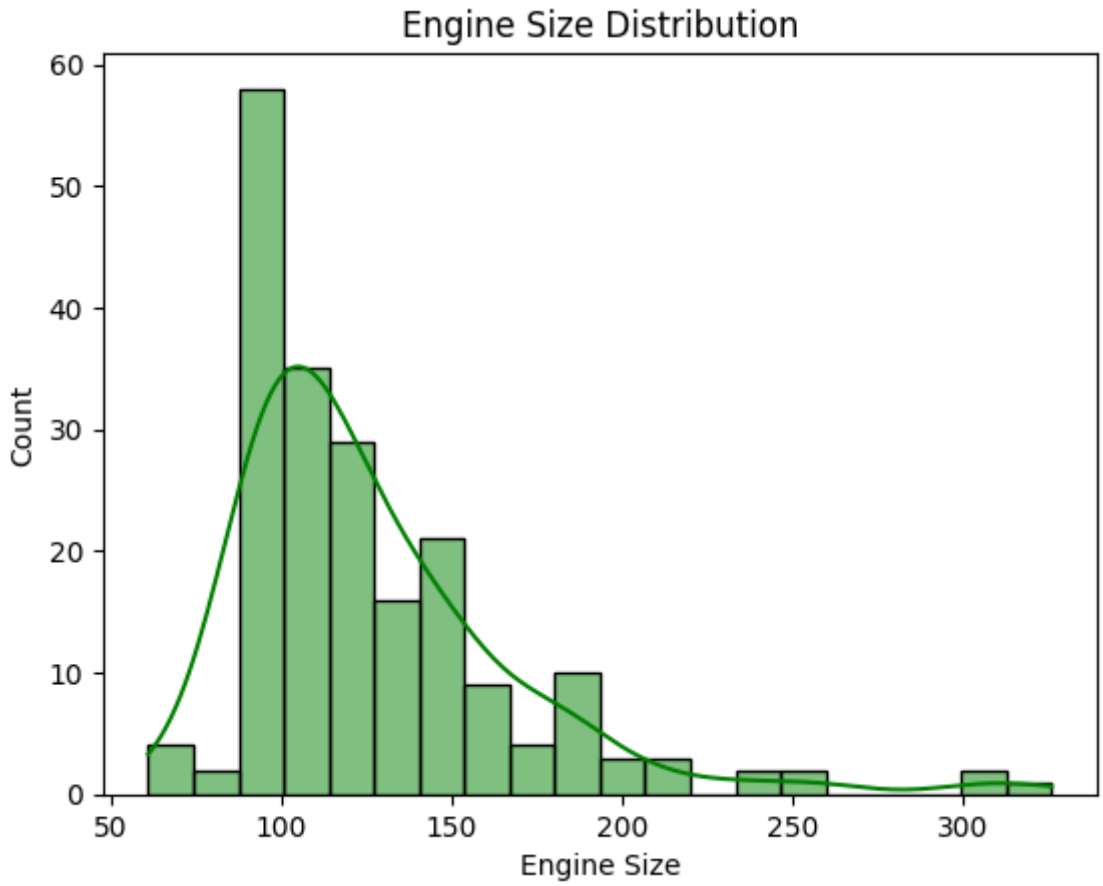
- This suggests majority of the cars are not fuel-efficient while driving in the highway , this is supported by the negatively skewed Boxplot and distribution line.

Distribution of Engine Size:

We will be able to find out the sizes of the engines in the Dataset, this will indicate if most of the cars in the Dataset have big or small engines

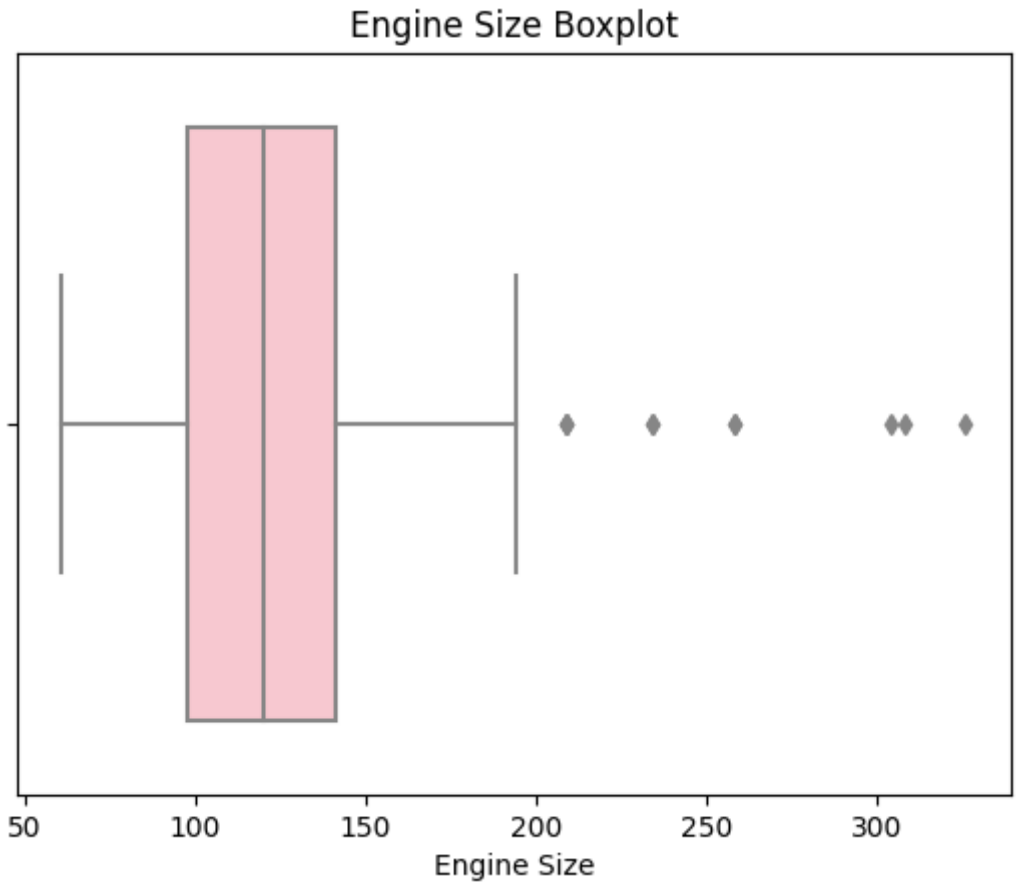
In [17]:

```
# Create a histogram using Seaborn
sns.histplot(data=df, x="engine-size", bins=20, kde=True, color="green")
plt.title("Engine Size Distribution")
plt.xlabel("Engine Size")
plt.ylabel("Count")
plt.show()
```



In [18]:

```
# Create a boxplot using Seaborn
sns.boxplot(x=df["engine-size"], color='pink')
plt.title("Engine Size Boxplot")
plt.xlabel("Engine Size")
plt.show()
```



Findings:

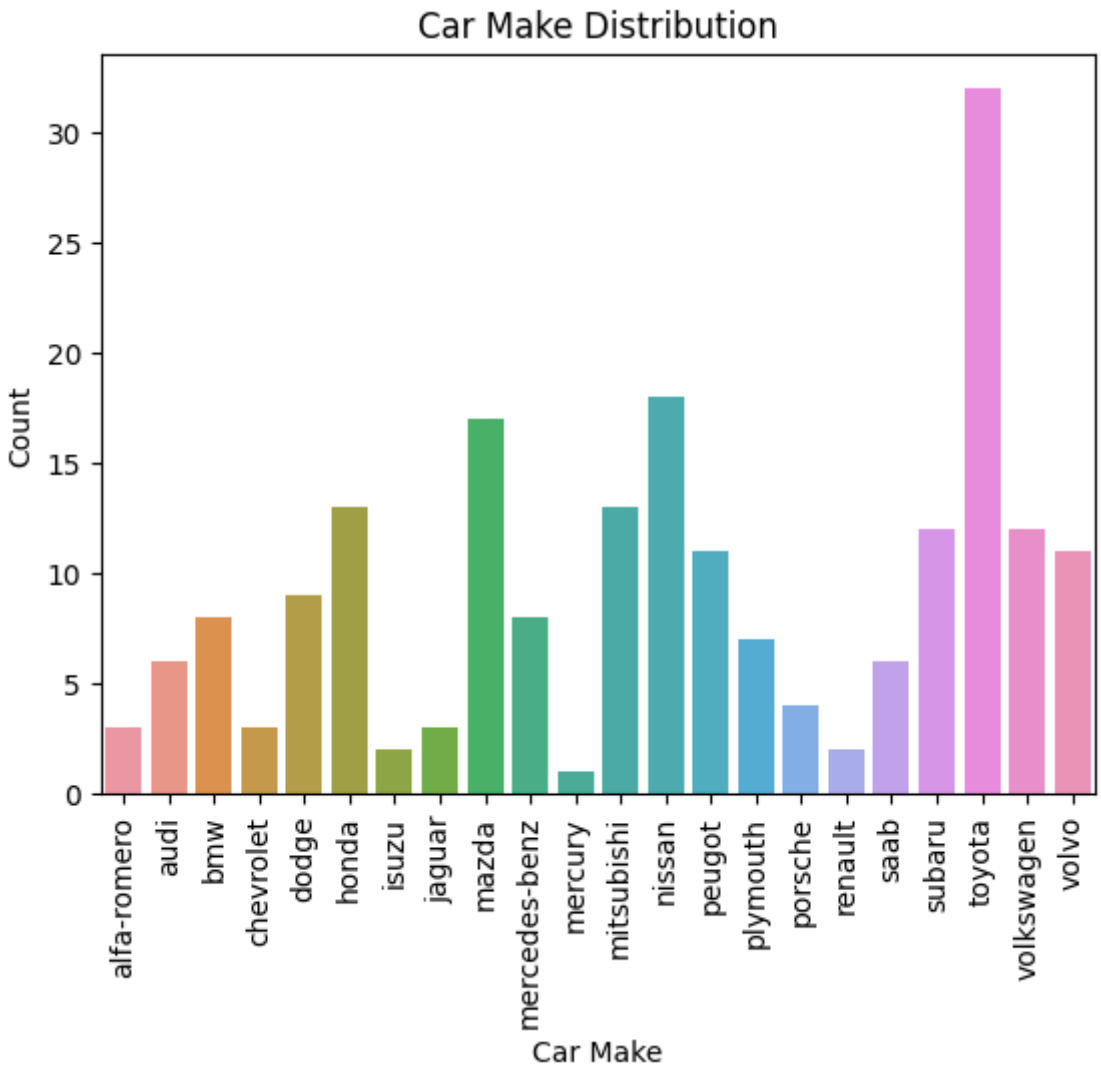
- With most of the cars below average, it simply means most of the cars in the Dataset have a smaller engine. However with the present of the outliers, this means that there are some cars with very large engines not representative of the majority of the cars in the in the dataset

Car Make distribution:

we will be able to see which car make are in the dataset and which one is more frequent.

In [19]:

```
# Create a count plot for Car Make
sns.countplot(data=df, x="make")
plt.title("Car Make Distribution")
plt.xlabel("Car Make")
plt.ylabel("Count")
plt.xticks(rotation=90)
plt.show()
```



Findings:

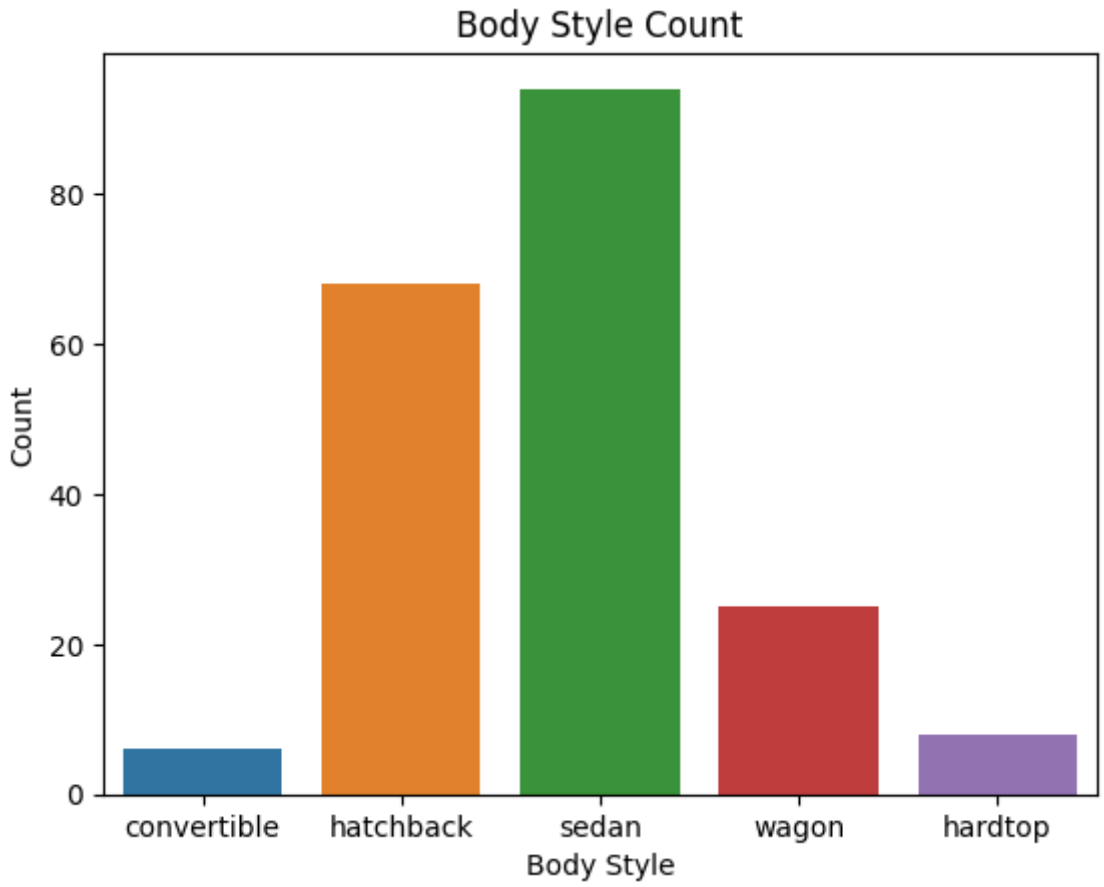
Toyota is the most Frequent Car make in the dataset, so most of the cars in the dataset are manufactured by Toyota.

Body style distribution:

To identify the body styles present in the Dataset and which one is the most frequent one.

In [20]:

```
# create a count plot for Body style
sns.countplot(data=df, x='body-style')
plt.title('Body Style Count')
plt.xlabel('Body Style')
plt.ylabel('Count')
plt.show()
```




```
Findings:

sedan body styles are the most frequent in the dataset.
```

2. Bivariate Analysis

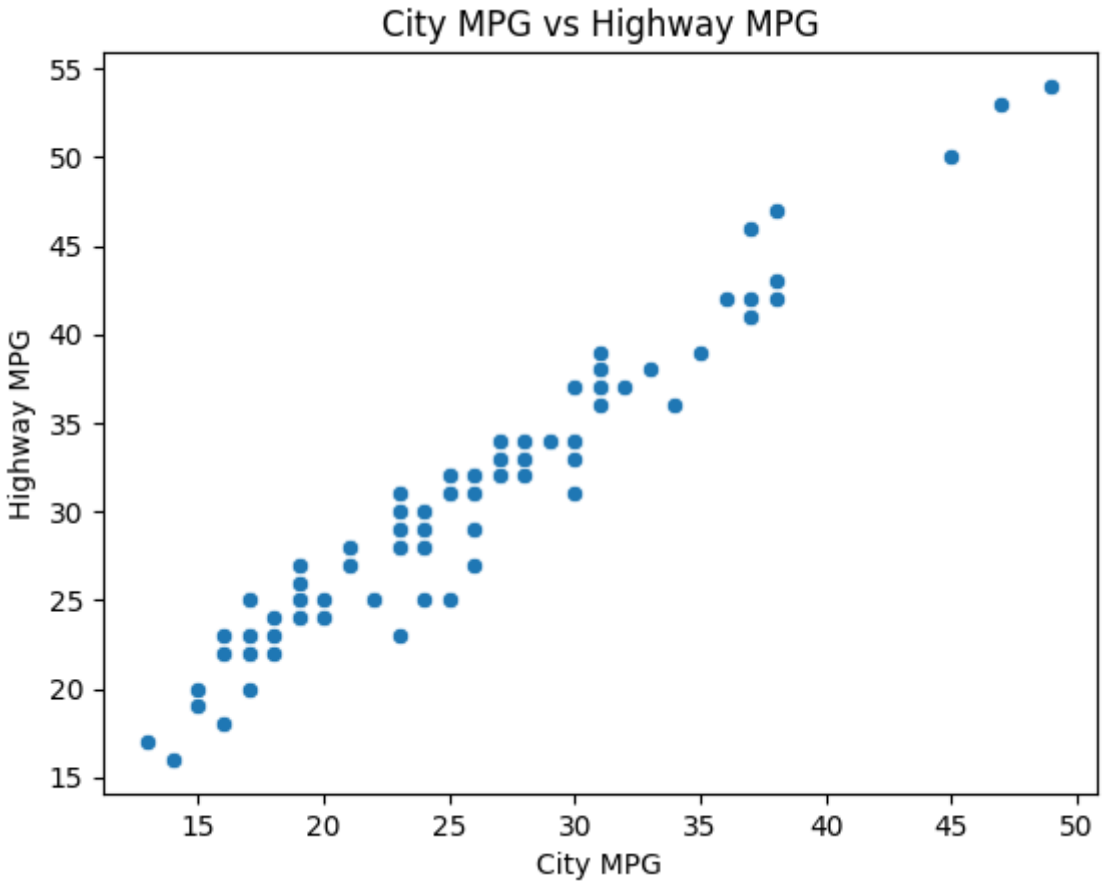
```
Exploring Relationships between two variables
```

```
Scatterplot of city mpg vs highway mpg:

we will see the relationship between these two variables, this plot will show if the
is any correlation present. This two variables are used to measure fuel efficiency
and will be vital in studying the factors that affect fuel efficiency.
```

In [21]:

```
# Plot a scatterplot of city mpg vs highway mpg
sns.scatterplot(x='city-mpg', y='highway-mpg', data=df)
plt.title('City MPG vs Highway MPG')
plt.xlabel('City MPG')
plt.ylabel('Highway MPG')
plt.show()
```



Findings:

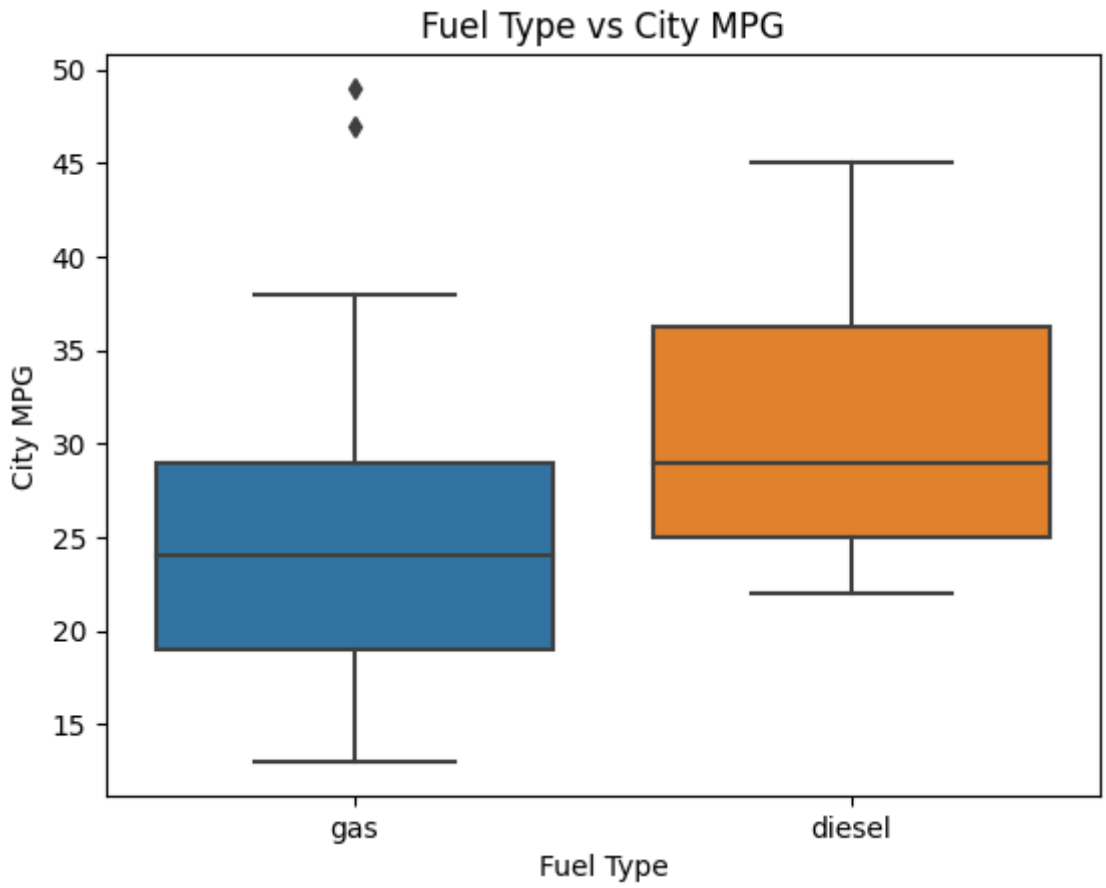
The is a positive correlation between the two variables , meaning they share a positive relationship.So because of that we can only visualise one MPG which at this case will be City Mpg

City Mpg vs Fuel Type:

We will be able to see which cars are more fuel efficient depending on the type of Fuel Type they use

In [22]:

```
# Create a boxplot for City vs Fuel Type
sns.boxplot(data=df, x="fuel-type", y="city-mpg")
plt.title("Fuel Type vs City MPG")
plt.xlabel("Fuel Type")
plt.ylabel("City MPG")
plt.show()
```



Findings:

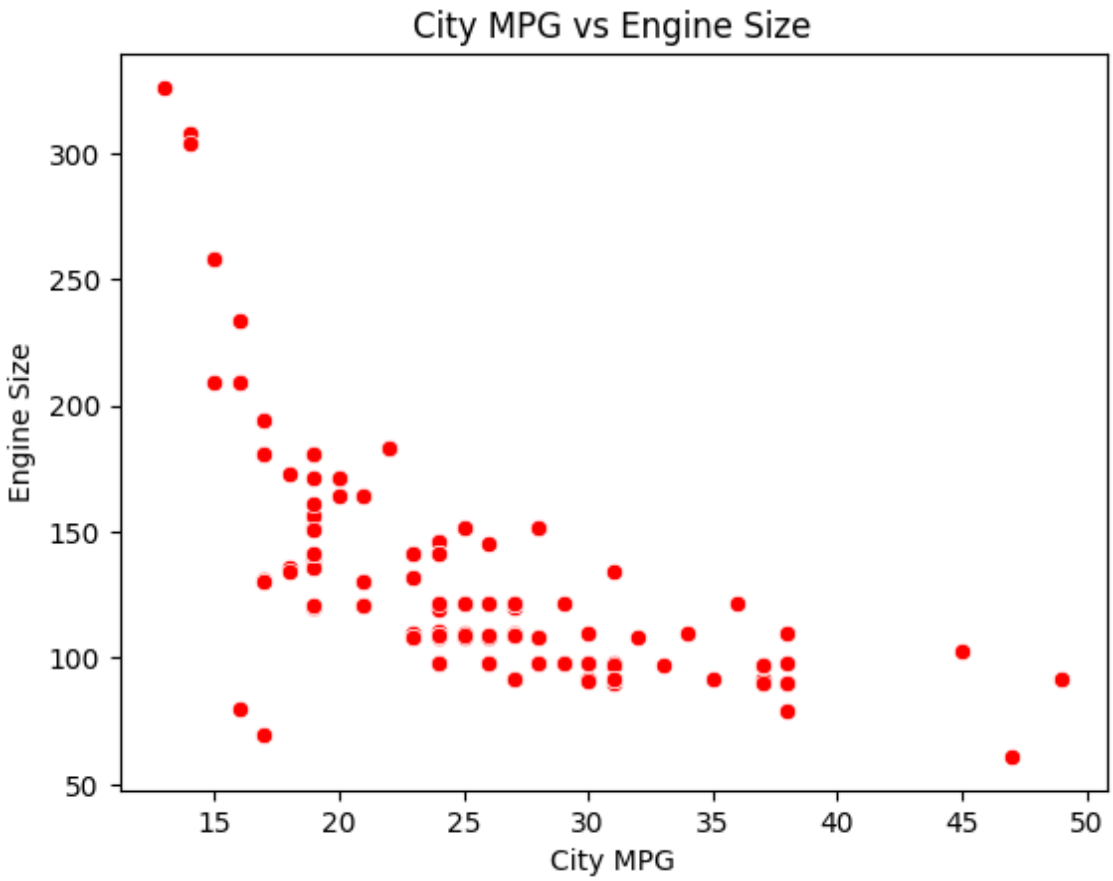
We Find that cars that use diesel are more fuel efficient than cars that use gas. It is also interesting to note that this plot still support the point that most of the cars in the dataset are more fuel efficient.

City MPG vs Engine Size:

We will identify the relationship between Engine size and fuel efficiency. We want to understand whether the engine has an impact on fuel effiency or not.

In [23]:

```
# Create a scatterplot for City MPG vs Engine Size
sns.scatterplot(data=df, x="city-mpg", y="engine-size", color ="red")
plt.title("City MPG vs Engine Size")
plt.xlabel("City MPG")
plt.ylabel("Engine Size")
plt.show()
```



In [24]:

```
# Check for the correlation coefficient
corr_coeff = df["city-mpg"].corr(df["engine-size"])

print(corr_coeff)
```

```
-0.6505459759740394
```

Findings:

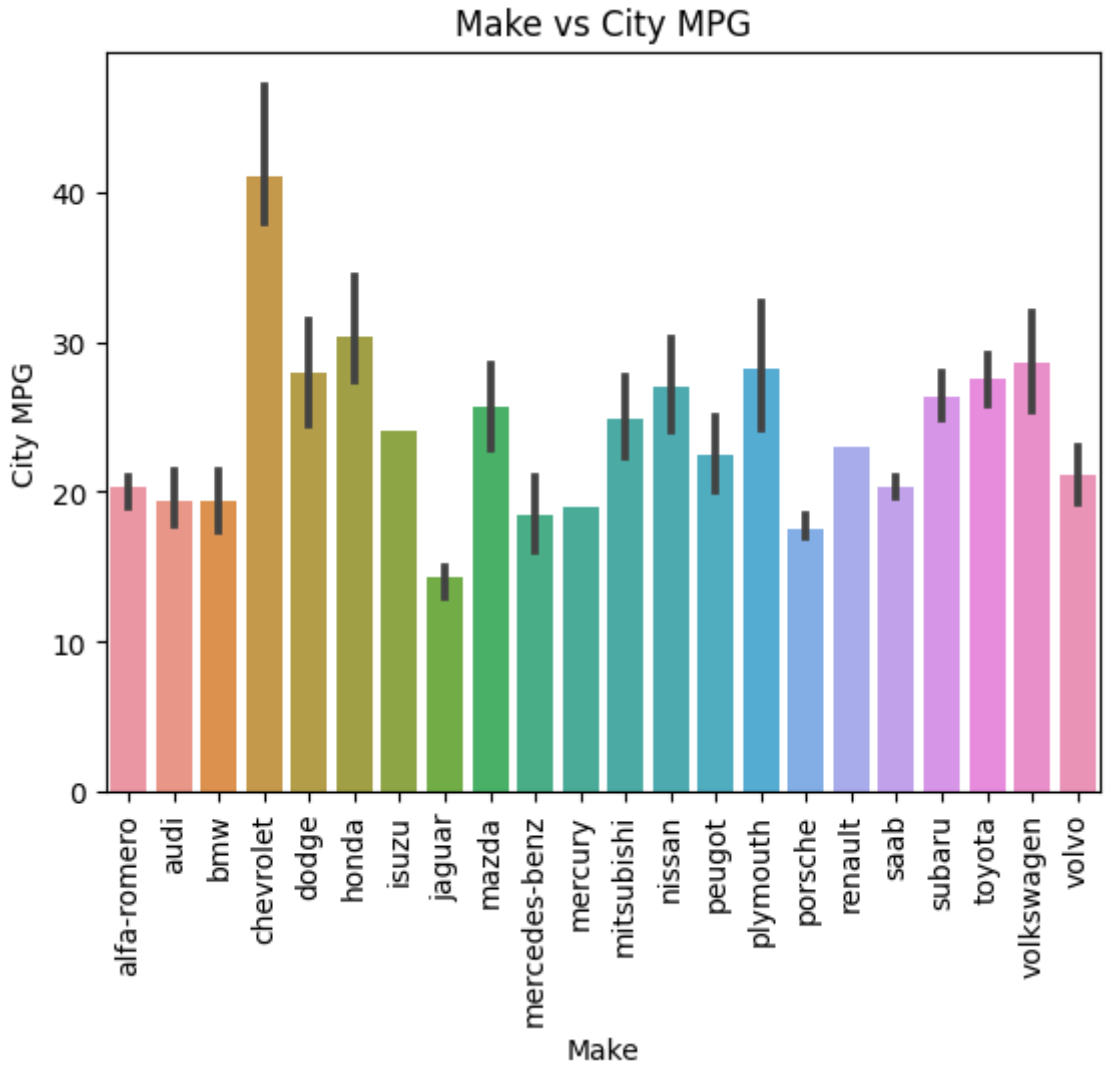
The is a strong negative relationship between the variables , this means that for every increase in engine size, the is a decline in fuel efficiency. So The size of a car's engine does impact fuel efficiency. Cars that have big engine sizes are less fuel efficient.

Make Vs City MPG:

We want to find out which car makes(brands) have higher fuel efficiency and which one has the least.

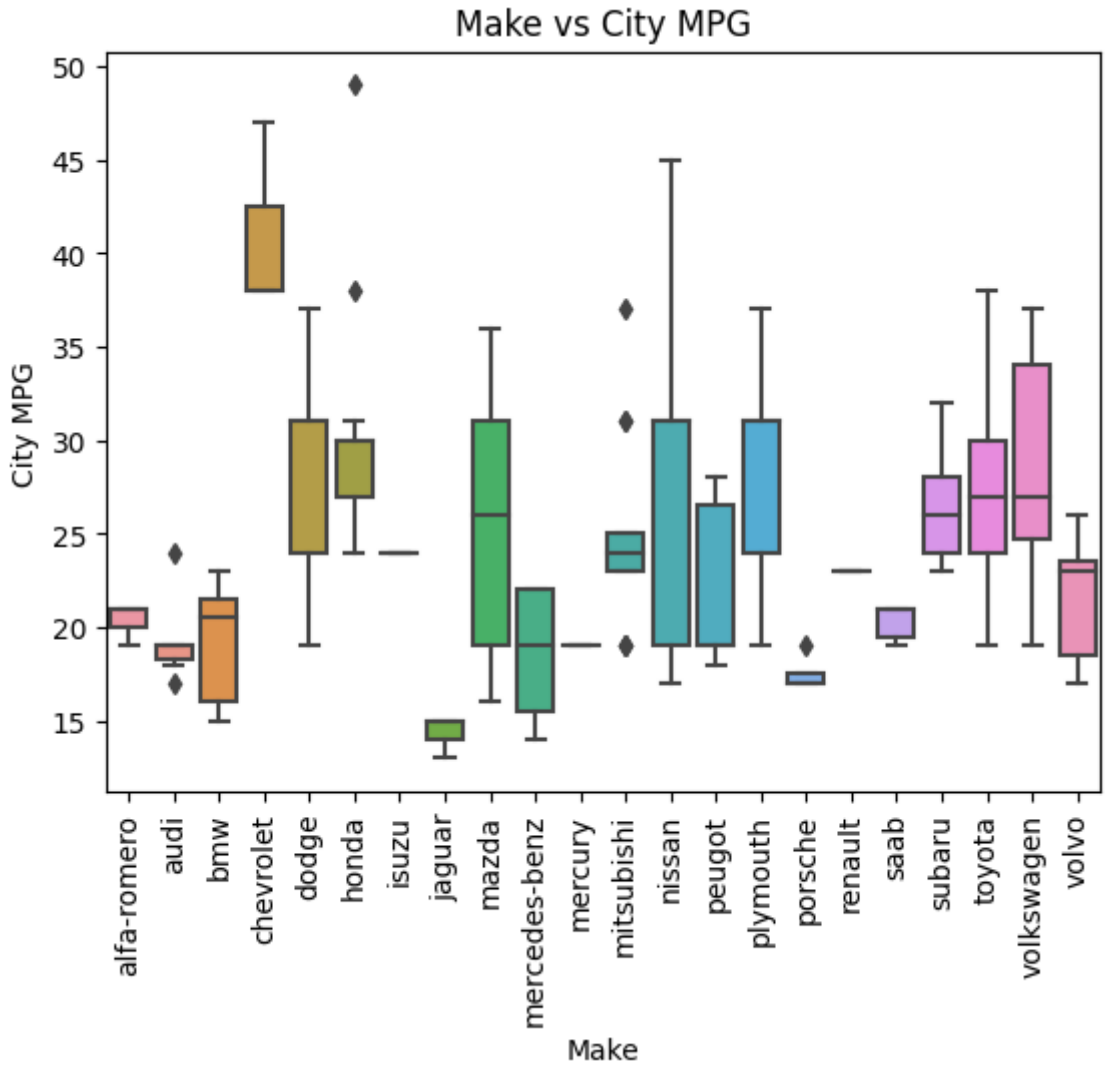
In [25]:

```
# Create a Barplot for Make Vs City MPG
sns.barplot(data=df, x="make", y="city-mpg")
plt.title("Make vs City MPG")
plt.xlabel("Make")
plt.ylabel("City MPG")
plt.xticks(rotation=90)
plt.show()
```



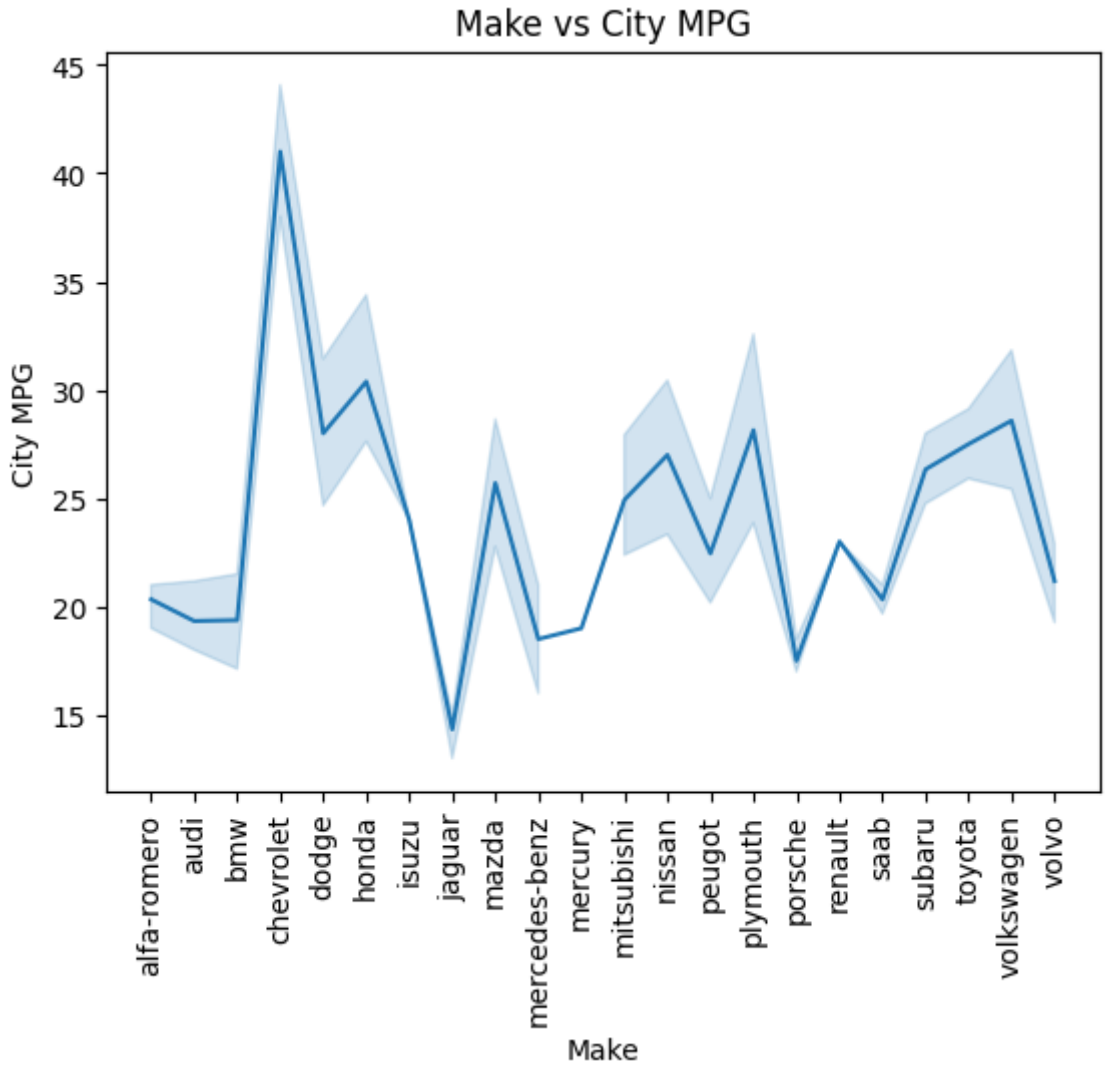
In [26]:

```
#Create a boxplot for Make Vs City MPG
sns.boxplot(data=df, x="make", y="city-mpg")
plt.title("Make vs City MPG")
plt.xlabel("Make")
plt.ylabel("City MPG")
plt.xticks(rotation=90)
plt.show()
```



In [27]:

```
# Create a lineplot for Make Vs City MPG
sns.lineplot(data=df, x="make", y="city-mpg")
plt.title("Make vs City MPG")
plt.xlabel("Make")
plt.ylabel("City MPG")
plt.xticks(rotation=90)
plt.show()
```



Findings:

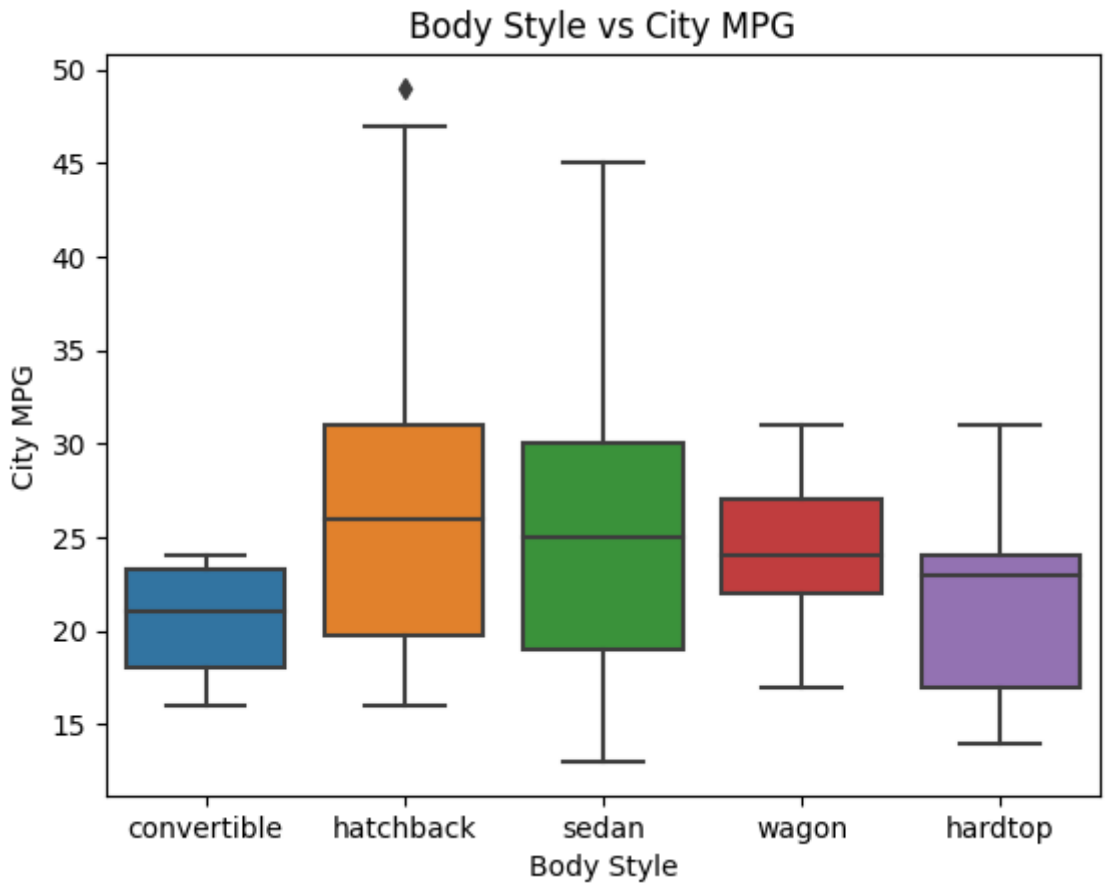
Chevrolet is the most fuel effiecient car make in the dataset, followed by Isuzu with jaguar being the least fuel efficient car in the Dataset.

Body style vs City MPG:

To find out which Body style has or is associated with high fuel efficiency.

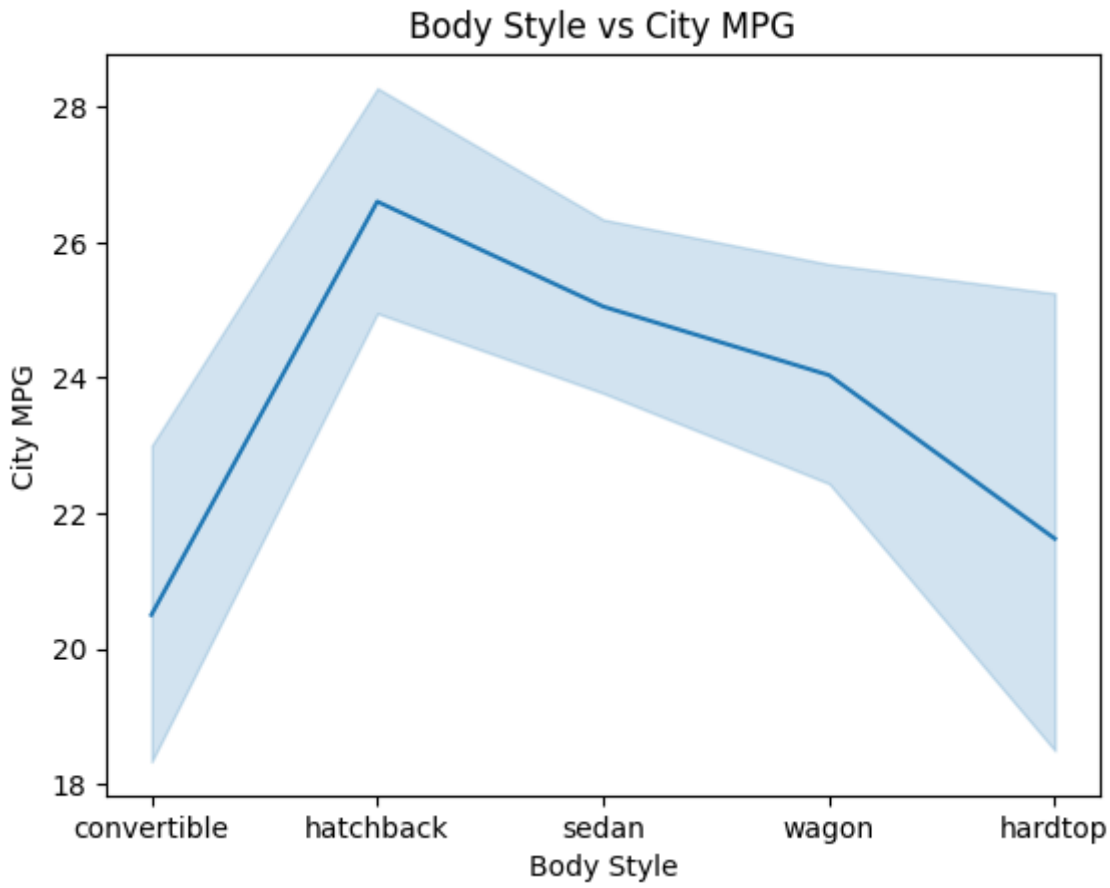
In [28]:

```
# Create a boxplot for Body style vs City MPG
sns.boxplot(data=df, x="body-style", y="city-mpg")
plt.title("Body Style vs City MPG")
plt.xlabel("Body Style")
plt.ylabel("City MPG")
plt.show()
```



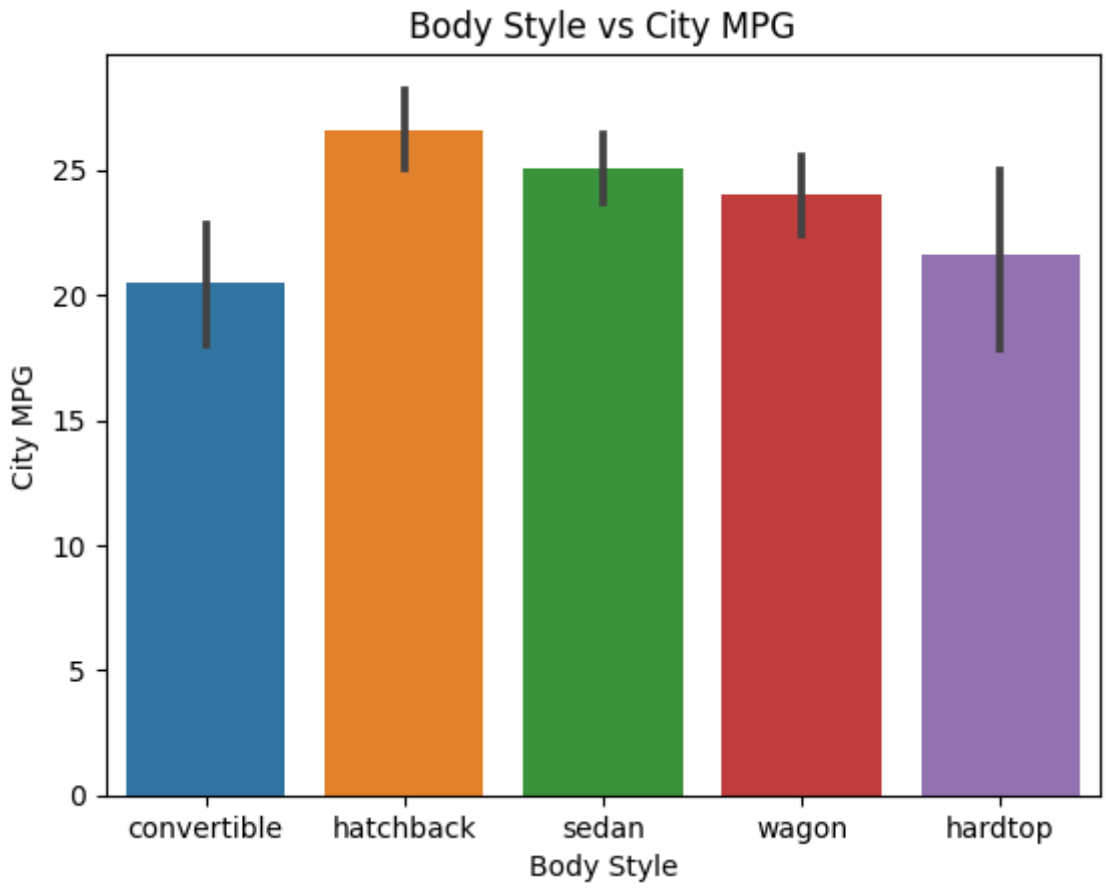
In [29]:

```
# Create a lineplot for Body style vs City MPG
sns.lineplot(data=df, x="body-style", y="city-mpg")
plt.title("Body Style vs City MPG")
plt.xlabel("Body Style")
plt.ylabel("City MPG")
plt.show()
```



In [30]:

```
# Create a Barplot for Body style vs City MPG
sns.barplot(data=df, x="body-style", y="city-mpg")
plt.title("Body Style vs City MPG")
plt.xlabel("Body Style")
plt.ylabel("City MPG")
plt.show()
```



Findings:

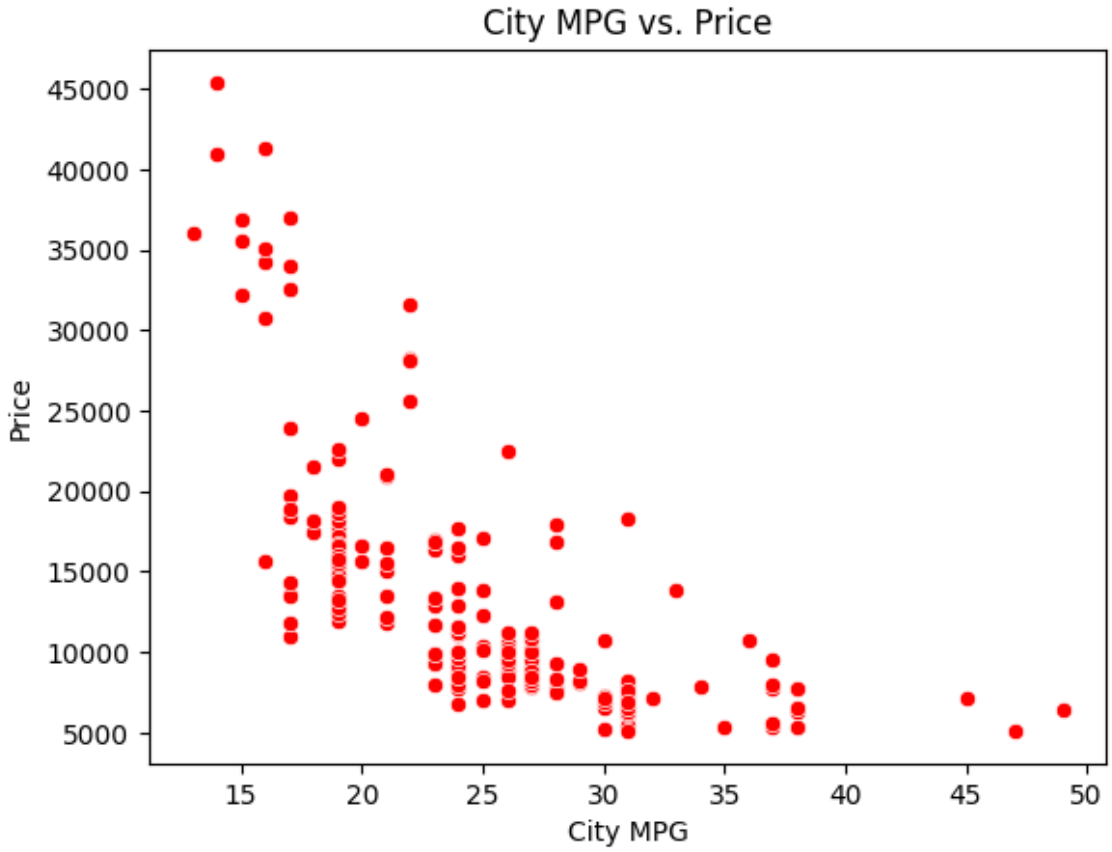
Cars with hatchback and sedan body styles are more fuel efficient than those with hardtop or convertible.

City vs Price:

we want to figure out the relationship between fuel efficieny and the price of a car.

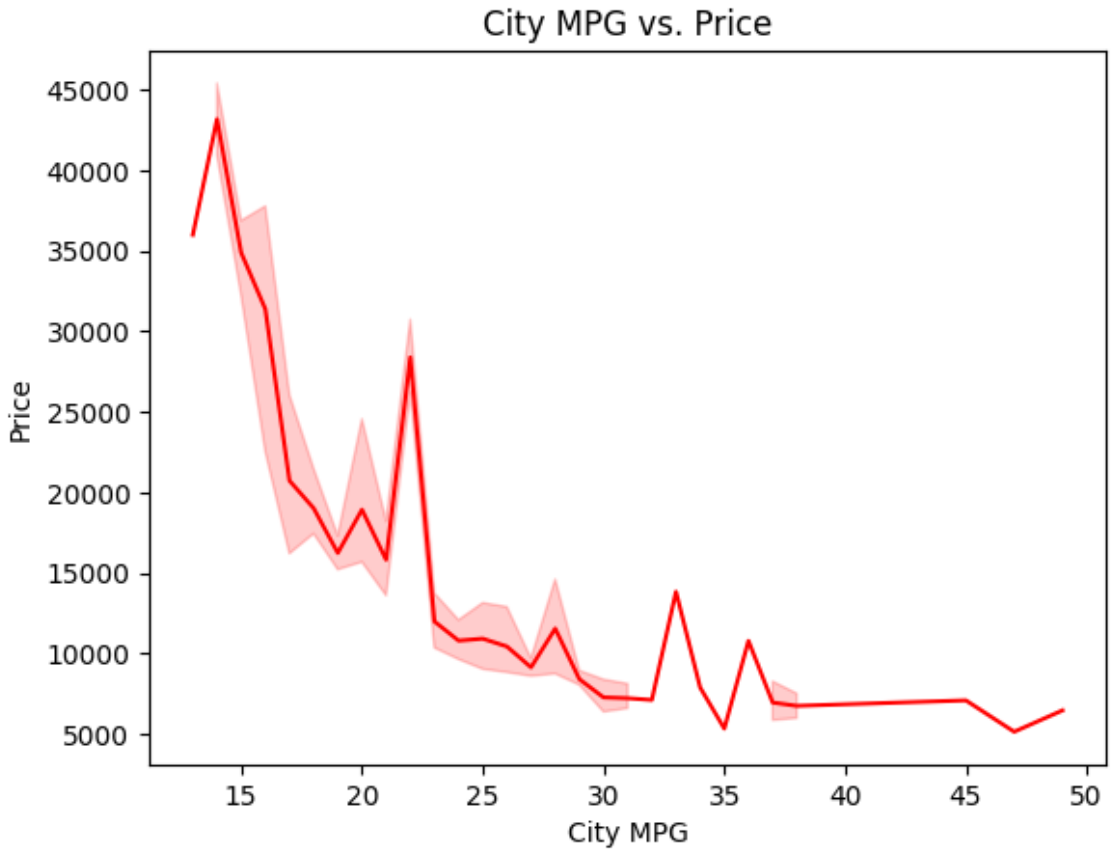
In [31]:

```
# Create a scatterplot for city mpg vs price
sns.scatterplot(data=df, x="city-mpg", y="price", color="red")
plt.title("City MPG vs. Price")
plt.xlabel("City MPG")
plt.ylabel("Price")
plt.show()
```



In [32]:

```
# Create a lineplot for city mpg vs price
sns.lineplot(data=df, x="city-mpg", y="price", color="red")
plt.title("City MPG vs. Price")
plt.xlabel("City MPG")
plt.ylabel("Price")
plt.show()
```



In [33]:

```
# Check for the correlation coefficient
corr_coeff = df["city-mpg"].corr(df["price"])

print(corr_coeff)
```

```
-0.686571006784468
```

```
Findings:

Car price and fuel efficiency have a strong negative relationship, this means that
expensive cars are less fuel efficient than cars that are cheaper.
```

3.Story

The Automibile dataset shows the details of car types with purpose of understanding fuel efficiency and identify the factors that affect fuel efficiency. The scope of the EDA is to analyze the following columns: Fuel Type, city mpg, highway mpg, engine size, make, and body styles.

We found that the are relationships that exists between this variables in relation to fuel efficiency. We found that while Toyota has more cars in the dataset, Chevrolet is the most fuel-efficient car make in the dataset. sedan body styles are the most frequent in the dataset, however Cars with hatchback and sedan body styles are more fuel efficient than those with hardtop or convertible. We also discovered that although most of the cars in the Dataset are use gas, cars that use diesel are more fuel efficient.

The analysis also revealed that price and fuel efficiency share a strong negative relationship, in the sense that the higher the price of the car the less fuel efficient it is. Same with the relationship we discovered about Engine size and fuel efficiency, the bigger the engine size the less the fuel efficiency.

4. Conclusion

1. Which car manufacturers produce the most fuel-efficient cars?
2. How does the engine size affect fuel efficiency?
3. Is there a difference in fuel efficiency based on the type of fuel used (gas or diesel)?
4. How does the body style of a car impact its fuel efficiency?
5. Does the price of a car have any relationship with its fuel efficiency?

1. Chevrolet is the most fuel effiecient car make in the dataset, followed by Isuzu with jaguar being the least fuel efficient car in the Dataset. So Chevrolet and Isuzu produced the most fuel efficient cars

2. The is a strong negative relationship between the variables , this means that for every increase in engine size, the is a decline in fuel efficiency. So The size of a car's engine does impact fuel efficiency. Cars that have big engine sizes are less fuel efficient.

3.cars that use diesel are more fuel efficient than cars that use gas.

4.Cars with hatchback and sedan body styles are more fuel efficient than those with hardtop or convertible.

5.Car price and fuel efficiency have a strong negative relationship, this means that expensive cars are less fuel efficient than cars that are cheaper.